Available online on 15.02.2026 at http://ajprd.com

Asian Journal of Pharmaceutical Research and Development

Open Access to Pharmaceutical and Medical Research

Open Access

Review Article

Formulation Modeling and Machine Learning In Injectable Drug Product Development: A Review

Dr. Bodkhe Atul Arvind*

PAR Formulations Pvt. Ltd, Mumbai, Maharashtra, India.

A B S T R A C T

Injectable drug products play a vital role in modern medicine, providing efficient and precise delivery of therapeutics. The development and optimization of injectable formulations require extensive research and development, as well as a deep understanding of the underlying physicochemical properties of the drug and its interactions with excipients. In recent years, machine learning (ML) techniques have emerged as powerful tools for predicting and modeling various aspects of drug formulation, leading to enhanced efficiency and cost-effectiveness in the pharmaceutical industry. This review article provides an overview of the application of ML techniques in the formulation modeling of injectable drug products, highlighting their potential and challenges in improving drug development processes.

Keywords: Injectable drug, formulation modeling, Modern healthcare

A R T I C L E I N F O: Received 23 Oct. 2025; Review Complete 28 Dec. 2025; Accepted 18 Jan. 2026; Available online 15 Feb. 2026

Cite this article as:
Bodkhe A A, Formulation Modeling and Machine Learning In Injectable Drug Product Development: A Review, Asian Journal of Pharmaceutical Research and Development. 2026; 14(1):39-47, DOI: http://dx.doi.org/10.22270/ajprd.v14i1.1692

*Address for Correspondence:
Dr. Bodkhe Atul Arvind, PAR Formulations Pvt. Ltd, Mumbai, Maharashtra, India.

INTRODUCTION

Injectable drug products play a crucial role in modern healthcare. They are medications that are administered directly into the body through intravenous, intramuscular, or subcutaneous routes. These products are typically in liquid form and are contained in vials, ampules, prefilled syringes, or infusion bags.

Below some key reasons why injectable drug products are important:

1. Efficient and Rapid Delivery: Injectable drugs allow for precise dosing and rapid delivery of medications into the bloodstream, bypassing the digestive system. This is especially critical in emergency situations or when immediate therapeutic effects are required[1].
2. Higher Bioavailability: Injectables often have higher bioavailability compared to oral medications. They enter the bloodstream directly, avoiding the first-pass metabolism that occurs when drugs are taken orally. As a result, injectables can achieve more reliable and predictable drug levels in the body.
3. Controlled Drug Delivery: Injectable drug products can be formulated to release medication slowly over a specified period, providing a controlled release profile. This enables extended therapeutic effects, reduces the frequency of administration, and improves patient compliance[2].
4. Administration in Critical Situations: In certain medical conditions or emergencies, patients may not be able to take medications orally due to reduced consciousness, vomiting, or gastrointestinal issues. Injectable drugs are vital in these cases as they can be administered by healthcare professionals directly into the patient's system[3].
5. Administration of Large Molecules: Some drugs, such as certain biologics, peptides, or proteins, cannot be effectively delivered orally as they may be degraded by digestive enzymes. Injectable formulations enable the administration of these complex molecules, allowing for their therapeutic benefits to be realized.
6. Flexibility in Dosing: Injectable drugs offer flexibility in dosing, allowing healthcare providers to tailor the dosage to individual patient needs. This is particularly important in cases where precise titration or adjustment of medication is required, such as in critical care settings or oncology.
7. Parenteral Nutrition: Injectable drug products are crucial in providing essential nutrition to patients who are unable

to consume food orally or have impaired gastrointestinal function. Intravenous administration of parenteral nutrition solutions ensures that patients receive the necessary nutrients for their overall well-being.

8. Disease Management: Many chronic diseases require long-term treatment and management. Injectable medications provide a reliable and effective means of administering therapies for conditions such as diabetes, rheumatoid arthritis, multiple sclerosis, and cancer.
9. Increased Drug Stability: Some drugs may be unstable or easily degraded when exposed to the acidic environment of the stomach or liver enzymes. Injectable formulations can protect these drugs from degradation, ensuring their stability and efficacy[4].

Need for efficient formulation development

Efficient formulation development is essential in the pharmaceutical industry for several reasons. Below are some key points, highlighting the need for efficient formulation development:

Optimal Drug Delivery: Formulation development aims to optimize drug delivery systems to ensure the efficient and effective delivery of medications to the target site in the body. This includes considerations of stability, bioavailability, solubility, release rate, and compatibility with the intended route of administration[5].

Patient Compliance: Developing formulations that are convenient, easy to administer, and well-tolerated by patients can significantly improve compliance. Formulations that are designed to minimize side effects, reduce dosing frequency, or provide patient-friendly administration routes (such as oral solid dosage forms or transdermal patches) enhance patient adherence to prescribed therapies.

Enhanced Drug Stability: Proper formulation development can improve the stability of drugs, preventing degradation, and maintaining their potency over time. This is particularly important for drugs that are prone to degradation in certain environmental conditions, such as temperature or humidity. Stable formulations ensure that the drugs retain their efficacy throughout their shelf life.

Improved Bioavailability: Bioavailability refers to the proportion of a drug that enters the systemic circulation and becomes available at the target site. Formulation development strategies can enhance drug solubility, dissolution rate, and permeability, thereby improving the bioavailability of poorly soluble or poorly absorbed drugs. This is crucial for achieving therapeutic efficacy and optimizing drug utilization.

Tailored Release Profiles: Some drugs require specific release profiles to achieve desired therapeutic effects. Formulation development enables the design of controlled-release or extended-release formulations that release drugs over a specific period, reducing the frequency of dosing and improving patient convenience. Tailored release profiles can also help optimize drug concentrations in the body, minimize side effects, and maintain therapeutic efficacy[6].

Compatibility and Safety: Formulation development involves considering the compatibility of drug substances with excipients and packaging materials. Compatibility studies help identify any potential interactions that may affect drug stability, efficacy, or safety. Efficient formulation development ensures that the selected excipients and packaging materials are suitable and do not compromise drug quality or patient safety[7].

Intellectual Property Protection: Developing efficient formulations can provide pharmaceutical companies with a competitive edge and intellectual property protection. Patents can be obtained for novel formulations or drug delivery systems, allowing companies to commercialize their products exclusively for a specific period and recoup their investments in research and development[8].

Efficient formulation development is a complex and iterative process that involves understanding the physicochemical properties of drugs, selecting appropriate excipients, conducting preformulation studies, and performing formulation optimization and stability testing. It requires collaboration between various disciplines, including pharmaceutical scientists, chemists, pharmacologists, and regulatory experts, to ensure the successful development of safe, effective, and commercially viable drug products.

Role of machine learning in drug development

Machine learning plays an increasingly important role in various aspects of drug development. Here are some key areas where machine learning is utilized:

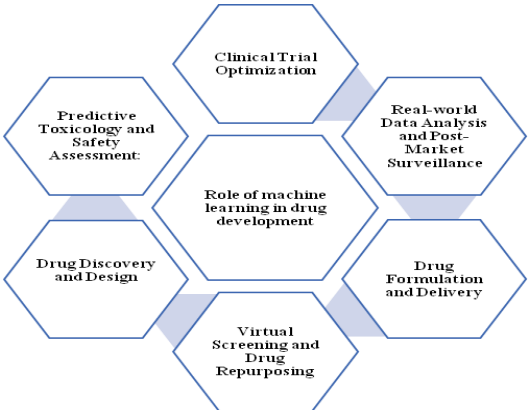

Figure 1: Role of machine learning in drug development

Machine learning algorithms are used to analyze vast amounts of biological and chemical data to identify potential drug targets, predict the activity of molecules, and design novel drug candidates. These algorithms can help optimize the drug discovery process by identifying patterns, predicting properties, and suggesting new molecular structures with desired properties[9].

Machine learning models can be trained on known drug-target interactions to predict new potential drug-target interactions. This enables virtual screening of large compound libraries to identify promising drug candidates. Additionally, machine learning techniques can aid in drug repurposing, identifying existing drugs that may have potential therapeutic uses beyond their original indications[10].

Machine learning algorithms are utilized to predict the toxicity and safety profiles of drug candidates. By analyzing large datasets of chemical structures and toxicity information, models can identify patterns and correlations to predict potential adverse effects. This helps prioritize compounds with favorable safety profiles and reduces the need for extensive animal testing[11].

Machine learning can optimize the design and execution of clinical trials by analyzing various data sources, including patient demographics, genetic information, biomarkers, and clinical outcomes. Predictive models can identify patient populations most likely to respond to a treatment, optimize dosage regimens, and predict potential adverse events. This aids in the efficient allocation of resources and improves the success rate of clinical trials[12].

Machine learning techniques can analyze real-world data from electronic health records, patient registries, and post-market surveillance databases to identify patterns and associations between drug use and outcomes. This helps in monitoring drug safety, detecting adverse events, and understanding the effectiveness of treatments in real-world settings[13].

Machine learning can assist in the optimization of drug formulations and delivery systems. Models can predict solubility, stability, and release profiles, aiding in formulation design. Additionally, machine learning algorithms can optimize drug delivery systems, such as nanoparticle-based drug carriers, to improve drug efficacy and target specific tissues or cells[14].

These are just a few examples of how machine learning is revolutionizing drug development. The integration of artificial intelligence and machine learning techniques has the potential to accelerate the discovery and development of new therapies, optimize treatment strategies, and improve patient outcomes. This article is focused on Machine Learning Techniques used in formulation development of injectable drug products[15].

Machine Learning Techniques in Formulation Modeling

Machine learning techniques are commonly used in formulation modeling to predict and optimize the properties and performance of formulations. Here are some popular machine learning techniques used in formulation modeling:

Regression Analysis: Regression analysis is commonly used to predict the properties of a formulation based on input variables such as composition, process conditions, and raw materials. Techniques like linear regression, polynomial regression, and support vector regression can be employed to develop predictive models.

Decision Trees: Decision trees are versatile and easy-to-understand models used for formulation modeling. They can handle both categorical and continuous input variables and provide insights into the relationship between the input variables and formulation properties. Decision tree-based algorithms like Random Forest and Gradient Boosting are often used for improved accuracy.

Artificial Neural Networks (ANNs): ANNs are powerful models that can capture complex relationships in formulation modeling. They consist of interconnected nodes (neurons) organized in layers, and each node performs a mathematical operation. ANNs can learn from data and make predictions based on learned patterns. Techniques like feed-forward neural networks, convolutional neural networks (CNNs), and recurrent neural networks (RNNs) can be used in formulation modeling tasks.

Support Vector Machines (SVMs): SVMs are effective models for both classification and regression problems. They map input variables to a higher-dimensional space and find the optimal hyperplane that separates data points or predicts a continuous output variable. SVMs are particularly useful when dealing with small datasets or when non-linear relationships exist in the formulation data.

Genetic Algorithms (GAs): GAs are optimization techniques inspired by natural evolution. They can be used to search for an optimal formulation composition or process conditions by iteratively evolving a population of potential solutions. GAs employ selection, crossover, and mutation operations to generate new solutions and improve the fitness of the population over generations.

Bayesian Networks: Bayesian networks are probabilistic graphical models that represent the probabilistic relationships between variables. They can be used to model the dependencies between formulation variables and predict properties or identify optimal conditions. Bayesian networks are particularly useful when dealing with uncertainty and limited data.

Principal Component Analysis (PCA): PCA is a dimensionality reduction technique that can be used to analyze and visualize the relationships between variables in a formulation dataset. It helps in identifying the most important variables and reducing the complexity of the modeling task.

These are just a few examples of machine learning techniques used in formulation modeling. The choice of technique depends on the specific problem, available data, and desired outcomes. Additionally, it's important to note that domain expertise and feature engineering play a crucial role in successful formulation modeling, as they help in selecting relevant variables and transforming raw data into meaningful features for the machine learning algorithms.

Overview of machine learning algorithms and approaches[16-25]

Machine learning algorithms and approaches are a broad and diverse set of techniques used to enable computers to learn from data and make predictions or decisions without being explicitly programmed. Here's an overview of some popular machine learning algorithms and approaches summarize in table below along with uses, advantages and disadvantages.

Table1: Machine learning algorithms summary

Category	Algorithms	Uses	Advantages	Disadvantages
Supervised Learning	Linear Regression	Predicting continuous output values	Simple and interpretable, computationally efficient	Assumes a linear relationship, sensitive to outliers
	Logistic Regression	Binary classification problems	Probabilistic interpretation, handles categorical features	Assumes a linear decision boundary, may suffer from overfitting
	Support Vector Machines (SVM)	Data classification, regression, outlier detection	Effective in high-dimensional spaces, works well with outliers	Computationally expensive for large datasets
	Decision Trees	Classification, regression	Easy to understand and interpret, handles missing values	Prone to overfitting, can create complex trees
	Random Forest	Classification, regression, feature selection	Robust to overfitting, handles high-dimensional data	Difficult to interpret the combined decision of multiple trees
	Gradient Boosting	Classification, regression	High predictive accuracy, handles complex interactions	Sensitive to noise and outliers
Unsupervised Learning	Clustering	Grouping similar data points	Discover hidden structures, no labeled data required	Requires determining the optimal number of clusters
	K-Means Clustering	Image segmentation, customer segmentation	Simple and efficient, scales well to large datasets	Sensitive to initialization and outliers
	Hierarchical Clustering	Taxonomy creation, gene expression analysis	Captures hierarchical relationships in the data	Computationally expensive for large datasets
	Dimensionality Reduction	Feature selection, visualization	Reduces data dimensionality, removes irrelevant features	May result in loss of information
	Principal Component Analysis (PCA)	Data compression, visualization	Captures maximum variance, reduces multicollinearity	Assumes linear relationships, may not preserve local structure
	t-SNE (t-Distributed Stochastic Neighbor Embedding)	Visualizing high-dimensional data	Preserves local similarities, reveals clusters and outliers	Computationally expensive for large datasets
	Association Rule Learning	Market basket analysis, recommendation systems	Discovers interesting relationships among items	High-dimensional data may result in a large number of rules
Reinforcement Learning	Q-Learning	Game playing, robot control	Learns optimal actions, handles complex environments	Sensitive to exploration-exploitation trade-off
	Deep Q-Networks (DQN)	Atari game playing, robotic control	Handles high-dimensional state spaces, improves convergence	Requires significant computational resources
	Policy Gradient Methods	Robotics, recommendation systems	Handles continuous action spaces, enables policy optimization	High variance, may converge to suboptimal policies
Deep Learning	Convolutional Neural Networks (CNN)	Image classification, object detection	Captures spatial dependencies, achieves state-of-the-art results	Computationally expensive for large models
	Recurrent Neural Networks (RNN)	Speech recognition, language modeling	Handles sequential data, captures temporal dependencies	Prone to vanishing/exploding gradient problem
	Long Short-Term Memory (LSTM)	Sentiment analysis, time-series prediction	Captures long-term dependencies, mitigates vanishing gradients	Computationally expensive for long sequences
	Generative Adversarial Networks (GAN)	Image synthesis, data augmentation	Generates realistic data, enables unsupervised representation learning	Training instability, mode collapse

Data preprocessing and feature selection

Data preprocessing and feature selection are important steps in machine learning to prepare the data for model training and improve the performance and efficiency of the learning algorithms. Here's an overview of data preprocessing and feature selection techniques:

Data Preprocessing:

Data Cleaning: Handling missing values, outliers, and noisy data through techniques like imputation, deletion, or outlier treatment[26].

Data Transformation: Scaling and normalizing the data to ensure all features have similar ranges or distributions. Common techniques include min-max scaling, standardization, or logarithmic transformation[27].

Encoding Categorical Variables: Converting categorical variables into numerical representations that machine learning algorithms can handle. Techniques include one-hot encoding, ordinal encoding, or target encoding[28].

Handling Imbalanced Data: Addressing class imbalance in the dataset by techniques such as oversampling, undersampling, or synthetic data generation (e.g., SMOTE)[29].

Feature Engineering: Creating new features from existing ones to capture additional information or simplify the representation. This can include mathematical transformations, interaction terms, or domain-specific knowledge[30].

Feature Selection:

Filter Methods: Assessing the relevance of features based on statistical measures like correlation, mutual information, or chi-square tests. Features are ranked or selected based on their scores[31].

Wrapper Methods: Evaluating subsets of features by training and evaluating the model performance on different combinations. Techniques like forward selection, backward elimination, or recursive feature elimination (RFE) are used[32].

Embedded Methods: Incorporating feature selection within the learning algorithm itself. Techniques like L1 regularization (Lasso), decision tree-based feature importance, or coefficient magnitudes in linear models[33].

Dimensionality Reduction: Transforming the data into a lower-dimensional space while preserving important information. Techniques like Principal Component Analysis (PCA) or t-SNE reduce the number of features by capturing their variations or similarities[34].

It's important to note that the choice of preprocessing and feature selection techniques depends on the specific characteristics of the dataset, the machine learning problem, and the algorithms being used. It is often an iterative process, and domain knowledge plays a crucial role in making informed decisions. Additionally, it's recommended to evaluate the impact of these techniques on the model performance using appropriate evaluation metrics and cross-validation.

Regression models for predicting drug properties

Regression models are commonly used in pharmaceutical research and drug development to predict various drug properties and characteristics. Below is summary of popular regression models used for predicting drug properties[35-44].

Table 2: Regression models for predicting drug properties

Regression Model	Description	Advantages
Multiple Linear Regression	Models the linear relationship between multiple independent variables and a continuous dependent variable	Simple interpretation, handles multiple features
Support Vector Regression	Extends support vector machines for regression tasks, effective in handling non-linear relationships	Handles high-dimensional featurespaces, robust to outliers
Random Forest Regression	Ensemble learning technique combining multiple decision trees for regression tasks	Handles non-linear relationships, robust to outliers
Gradient Boosting Regression	Ensemble learning technique that builds an ensemble of weak regression models iteratively	High accuracy, handles complex relationships
Neural Network Regression	Utilizes interconnected nodes (neurons) organized in layers to capture complex patterns in the data	Handles complex relationships, suitable for deep learning
Bayesian Regression	Incorporates prior knowledge and uncertainty into the regression task using Bayesian inference	Useful for limited data or when prior knowledge is available
Gaussian Process Regression	Non-parametric regression method modeling the regression problem as a distribution over functions	Handles uncertainty, suitable for small datasets

These are some examples of regression models commonly used for predicting drug properties. The selection of the appropriate model depends on the specific problem, the nature of the data, the complexity of the relationships, and the availability of data. It's often beneficial to compare and evaluate multiple models to determine the best-performing one for a particular drug property prediction task.

Classification models for formulation optimization

Classification models are commonly used in formulation optimization to predict the optimal formulation conditions or to classify formulations into different categories based on their properties. Here are some popular classification models used for formulation optimization[45-54]

Table 3: Classification models are commonly used in formulation optimization

Classification Model	Description	Advantages
Logistic Regression	Models the relationship between independent variables and a binary or categorical dependent variable	Interpretable, handles binary and categorical outcomes
Decision Trees	Hierarchical tree-like structures that make decisions based on feature values	Easy to understand, handles complex relationships
Random Forest	Ensemble learning technique combining multiple decision trees for classification tasks	High accuracy, handles complex relationships, resistant to overfitting
Support Vector Machines	Constructs a hyperplane to separate data points into different classes	Effective in high-dimensional spaces, handles non-linear relationships
Naive Bayes	Probabilistic classification algorithm based on Bayes' theorem with the assumption of independence	Simple and computationally efficient, works well with categorical features
Neural Networks	Interconnected nodes (neurons) organized in layers to capture complex patterns in the data	Handles complex relationships, suitable for deep learning
Gradient Boosting	Ensemble learning technique building an ensemble of weak classification models iteratively	High accuracy, handles complex relationships
K-Nearest Neighbors (KNN)	Classifies data points based on majority class among their k-nearest neighbors	Simple and effective, handles non-linear decision boundaries

These are some examples of classification models commonly used for formulation optimization. The choice of the appropriate model depends on the specific problem, the nature of the data, the complexity of the relationships, and the availability of data. It is often beneficial to compare and evaluate multiple models to determine the best-performing one for a particular formulation optimization task.

Clustering techniques for identifying formulation trends

Clustering techniques are commonly used in formulation analysis to identify trends and group similar formulations together based on their properties. These techniques help in discovering patterns and relationships among formulations. Here are some popular clustering techniques used for identifying formulation trends[55-64].

Table 4: Clustering techniques

Clustering Technique	Description	Advantages
K-means Clustering	Partition data into K clusters by iteratively updating centroids based on proximity	Simple and efficient, scalable to large datasets, easy interpretation of results
Hierarchical Clustering	Build a hierarchical structure of clusters by merging or splitting clusters based on similarity/dissimilarity	No need to specify the number of clusters in advance, provides a visual representation of clustering structure
DBSCAN	Group data points based on density and separate regions with low-density	Does not require specifying the number of clusters, robust to outliers, can discover clusters of arbitrary shape
Gaussian Mixture Models	Model data as a mixture of Gaussian distributions and assign data points based on maximum likelihood	Allows for soft assignment, can capture complex distributions, flexible in handling different cluster shapes
Self-Organizing Maps (SOM)	Map high-dimensional data onto a lower-dimensional grid while preserving topological relationships	Effective in visualizing and interpreting complex data, preserves topology of data points
Affinity Propagation	Determine exemplars (representative points) for clusters by passing messages between data points	Does not require specifying the number of clusters, can handle large datasets, automatically adapts to data characteristics
Mean Shift Clustering	Identify clusters as regions of high data density by iteratively moving a kernel	Does not require specifying the number of clusters, can handle irregularly shaped clusters, robust to noise

These are some examples of clustering techniques commonly used for identifying formulation trends. The choice of the appropriate technique depends on the specific problem, the nature of the data, and the desired outcomes. It is often useful to try multiple clustering techniques and evaluate the results to determine the most meaningful and interpretable formulation trends.

Prediction of Physicochemical Properties[65-67]

Prediction of physicochemical properties is a crucial task in various fields, including drug discovery, material science,

environmental analysis, and chemical engineering. Several machine learning techniques can be employed for predicting physicochemical properties based on molecular or chemical descriptors. Below are some commonly used approaches:

Table 5: Prediction of Physicochemical Properties

Machine Learning Technique	Description	Advantages
Quantitative Structure-Property Relationship (QSPR)	Establishes a relationship between physicochemical property and molecular descriptors	Enables prediction of continuous properties, interpretable models, captures structure-property relationships
Quantitative Structure-Activity Relationship (QSAR)	Correlates molecular descriptors with biological activity or toxicity data	Useful in drug discovery and risk assessment, predicts activity/toxicity based on structure and physicochemical properties
Deep Learning	Utilizes deep neural networks to learn complex patterns and relationships in molecular data	Captures intricate features, handles large datasets, suitable for predicting bioactivity, toxicity, and other complex properties
Random Forest	Ensemble learning technique that combines multiple decision trees	Handles non-linear relationships, accommodates a large number of descriptors, robust to noise and outliers
Support Vector Machines (SVM)	Powerful machine learning algorithm for classification and regression tasks	Effective for predicting continuous values, robust to high-dimensional data, can handle non-linear relationships
Bayesian Models	Incorporates prior knowledge and handles uncertainty estimation using Bayesian inference	Provides a probabilistic framework, handles uncertainty, allows for prior knowledge integration
Cheminformatics Tools	Software packages or tools with pre-built models and functionalities for property prediction	User-friendly interface, readily available models, ease of use, can cover a wide range of property prediction tasks

These approaches are just a few examples of the techniques used for predicting physicochemical properties. The selection of the appropriate method depends on the specific property to be predicted, the available data, and the desired level of accuracy and interpretability. It is often beneficial to compare and evaluate multiple models to determine the best-performing approach for a particular prediction task.[68-71]

Table 6: Research work with use of ML techniques

Topic	Description	ML Techniques Used
Solubility Prediction Model	Utilize various machine learning algorithms for solubility prediction	Support Vector Machines, Random Forest, Neural Networks, Gradient Boosting, Gaussian Processes
pH Stability Prediction	Prediction of the stability of a formulation under different pH conditions	Support Vector Machines (SVM), Random Forest, Artificial Neural Networks (ANN), Gaussian Process Regression, Quantitative Structure-Property Relationship (QSPR), Quantitative Structure-Activity Relationship (QSAR)
Particle Size and Suspension Stability Modeling	Modeling and prediction of particle size and stability of suspensions or colloidal systems	Support Vector Regression (SVR), Random Forest, Artificial Neural Networks (ANN), Decision Trees, Gaussian Process Regression
Excipient Selection and Compatibility Studies	Selection of suitable excipients for a formulation and assessment of compatibility between excipients and active pharmaceutical ingredients	Bayesian Models, Decision Trees, Support Vector Machines (SVM), Random Forest, Naive Bayes, Logistic Regression, Gaussian Process Regression
Prediction of Excipient-Drug Interactions	Prediction of potential interactions between excipients and drugs based on their physicochemical properties	Support Vector Machines (SVM), Random Forest, Decision Trees, Gaussian Process Regression, Bayesian Models, Logistic Regression
Compatibility Assessment Using ML Techniques	Assessment of compatibility between different components in a formulation using machine learning techniques	Random Forest, Support Vector Machines (SVM), Decision Trees, Gaussian Process Regression, Bayesian Models, Logistic Regression
Formulation Optimization and Stability Prediction	Optimization of formulation components and prediction of formulation stability	Genetic Algorithms, Artificial Neural Networks (ANN), Support Vector Machines (SVM), Bayesian Models, Random Forest, Decision Trees, Gaussian Process Regression
Design of Experiments (DoE) Using ML	Application of machine learning techniques to optimize experimental designs for formulation development	Support Vector Machines (SVM), Random Forest, Decision Trees, Bayesian Models, Artificial Neural Networks (ANN), Gaussian Process Regression

Topic	Description	ML Techniques Used
Predictive Modeling of Formulation Stability	Development of predictive models to forecast formulation stability over time	Random Forest, Artificial Neural Networks (ANN), Support Vector Machines (SVM), Gaussian Process Regression
Real-Time Monitoring of Injectable Formulations	Real-time monitoring of critical parameters in injectable formulations using ML techniques	Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forest, Gaussian Process Regression

Challenges and Limitations

Interpretability and Transparency of ML Models.

ML models often lack interpretability and transparency, making it challenging to understand the reasoning behind the predictions.Techniques like SHAP (SHapley Additive exPlanations) values, feature importance analysis, and model-agnostic interpretability methods can be employed to enhance interpretability.

Regulatory Considerations and Validation Requirements

Regulatory agencies require validation of ML models for use in formulation and stability prediction tasks

Rigorous validation protocols, including model performance assessment, robustness testing, and external validation, must be followed to meet regulatory standards.

Future Perspectives

Integration of ML with other Computational Tools. ML techniques can be integrated with other computational tools, such as molecular docking or molecular dynamics simulations, for comprehensive formulation analysis.Combined approaches can provide a more holistic understanding of formulation behavior and properties.Use of ML in Personalized Medicine and Patient-Specific Formulations. ML can enable the development of personalized medicine by predicting formulation behavior and optimizing drug delivery for individual patients.

Conflicts of interest

CONCLUSION

This comprehensive review article provides an in-depth exploration of the application of machine learning techniques in the formulation modeling of injectable drug products. It covers various aspects, ranging from physicochemical property prediction to excipient selection, formulation optimization, and stability prediction. The article emphasizes the potential benefits and challenges associated with the implementation of ML in the pharmaceutical industry. Additionally, it discusses future perspectives and identifies areas for further research to enhance the integration of ML into the drug development process.

Overall, this review article serves as a valuable resource for researchers, scientists, and professionals in the pharmaceutical field, providing insights into the latest advancements and potential applications of machine learning in injectable drug product formulation modeling.

REFERENCES:

1. World Health Organization (WHO). (2010). WHO best practices for injections and related procedures toolkit. Retrieved from https://www.who.int/infection-prevention/tools/injections/injections/en/
2. Dua, P., Hawkins, E. G., & Lal, C. V. Injectable drug delivery systems: An overview. In Comprehensive Biotechnology (Second Edition) 2016; 2: 57-69. Elsevier. doi: 10.1016/B978-0-444-63428-3.00005-4
3. Cevher, E., &Sensoy, D. (Eds.). (2018). Injectable Drug Delivery Systems: From Concept to Clinical Practice. CRC Press.
4. Rathore, A. S., Pathak, S., & Vyas, S. P. (2019). Injectable drug delivery systems: An overview. In Drug Delivery Systems (pp. 1-25). Woodhead Publishing. doi: 10.1016/B978-0-08-102550-1.00001-5.
5. Bala, R., Pawar, P., & Khanna, S. (2017). Formulation and Development of Pharmaceutical Dosage Forms. In Encyclopedia of Pharmacy Practice and Clinical Pharmacy (pp. 72-90). Academic Press. doi: 10.1016/B978-0-12-812736-0.00009-0.
6. Rangaraj, N., & Reddy, L. H. Pharmaceutical formulation development: A quality by design approach. Journal of Pharmaceutical Investigation, 2014; 44(5):309-320. doi: 10.1007/s40005-014-0145-6
7. Rizvi, S. A. A., & Saleh, A. M. (2018). Applications of Quality by Design (QbD) for Developing Pharmaceutical Dosage Forms. In Quality by Design Approaches in Drug Delivery Systems (pp. 1-20). Springer. doi: 10.1007/978-3-319-66258-5_1
8. Shah, S., & Patel, M. (2019). Importance of preformulation in formulation development. In Preformulation in Solid Dosage Form Development (pp. 1-13). Elsevier. doi: 10.1016/B978-0-12-816806-6.00001-0.
9. Vippagunta, S. R., &Repka, M. A. (2018). Role of formulation development in drug discovery and development. In Pharmaceutical Formulation Development of Peptides and Proteins (pp. 1-12). CRC Press. doi: 10.1201/9781315119467-1.
10. Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P. & Xiao, C. (2018). Opportunities and obstacles for deep learning in biology and medicine. Journal of The Royal Society Interface, 15(141), 20170387. doi: 10.1098/rsif.2017.0387
11. Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., &Svetnik, V. (2015). Deep neural nets as a method for quantitative structure–activity relationships. Journal of Chemical Information and Modeling, 2015; 55(2), 263-274. doi: 10.1021/ci500747n.
12. Ramsundar, B., Liu, B., Wu, Z., Verras, A., Tudor, M., Sheridan, R. P., & Pande, V. Is multitask deep learning practical for pharma? Journal of Chemical Information and Modeling, 2017; 57(8):2068-2076. doi: 10.1021/acs.jcim.7b00146.
13. Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A. & Webster, D. R. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 2016; 316(22):2402-2410. doi: 10.1001/jama.2016.17216
14. Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2017). Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics,2017; 19(6):1236-1246. doi: 10.1093/bib/bbx044
15. Ramsundar, B., Eastman, P., Walters, P., Pande, V. S., &Leswing, K. (2019). Deep learning for the life sciences: Applying deep learning to genomics, microscopy, drug discovery, and more. O'Reilly Media.
16. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
17. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
18. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
19. Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.

20. Alpaydin, E. (2020). Introduction to Machine Learning (3rd ed.). The MIT Press.
21. Mitchell, T. (1997). Machine Learning. McGraw-Hill
22. Sutton, R. S., &Barto, A. G. (2018). Reinforcement Learning: An Introduction. MIT Press.
23. Chollet, F. (2017). Deep Learning with Python. Manning Publications.
24. Russell, S. J., & Norvig, P. (2016). Artificial Intelligence: A Modern Approach (3rd ed.). Pearson.
25. Marsland, S. (2015). Machine Learning: An Algorithmic Perspective (2nd ed.). CRC Press.
26. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
27. Guyon, I., &Elisseeff, A. An introduction to variable and feature selection. Journal of Machine Learning Research, 2003; 3:1157-1182.
28. Kotsiantis, S. B., Zaharakis, I. D., &Pintelas, P. E. Machine learning: A review of classification and combining techniques. Artificial Intelligence Review,2006; 26(3):159-190.
29. Brownlee, J. (2019). Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. Machine Learning Mastery.
30. Chandrashekar, G., &Sahin, F. A survey on feature selection methods. Computers & Electrical Engineering, 2014; 40(1):16-28.
31. Guyon, I., &Elisseeff, A. Feature selection with ensembles, artificial variables, and redundancy elimination. Journal of Machine Learning Research, 2006; 7:1293-1315.
32. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. &Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 2011; 12(10): 2825-2830.
33. Brownlee, J. (2018). Data Cleaning for Machine Learning: How to Automatically Clean Data. Machine Learning Mastery
34. Jolliffe, I. (2011). Principal Component Analysis (2nd ed.). Springer.
35. Tropsha, A., &Gramatica, P. (2003). QSAR in Drug Design and Toxicology: Historical Perspective and Recent Advances. Springer Science & Business Media.
36. Chen, Y., & Zou, P. Predicting Drug-Target Interactions from Chemical and Genomic Data with Network Fusion-based Models. Computational and Structural Biotechnology Journal, 2017; 15:378-384.
37. Zhang, G., & Xi, H. Predicting Drug-Target Interactions Using Deep Learning Models. Journal of Chemical Information and Modeling, 2019; 59(2), 615-624.
38. Golbraikh, A., &Tropsha, A. Beware of q2! Journal of Molecular Graphics and Modelling, 2002; 20(4):269-276.
39. Segler, M. H., & Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. Chemistry - A European Journal, 2017; 23(25), 5966-5971.
40. Sliwoski, G., et al.. Computational Methods in Drug Discovery. Pharmacological Reviews, 2014; 66(1); 334-395.
41. Cawley, G. C., & Talbot, N. L. C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. Journal of Machine Learning Research, 2010; 11:2079-2107.
42. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
43. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
44. Brownlee, J. (2016). Machine Learning Mastery with Python. Machine Learning Mastery.
45. Brownlee, J. (2019). Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models and Work Projects End-To-End. Machine Learning Mastery.
46. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
47. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer
48. Raschka, S., &Mirjalili, V. (2020). Python Machine Learning (3rd ed.). Packt Publishing.
49. James, G., Witten, D., Hastie, T., &Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. Springer.
50. Tan, P. N., Steinbach, M., & Kumar, V. (2018). Introduction to Data Mining (2nd ed.). Pearson.
51. Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.
52. Chollet, F. (2017). Deep Learning with Python. Manning Publications.
53. Flach, P. (2012). Machine Learning: The Art and Science of Algorithms that Make Sense of Data. Cambridge University Press
54. Caruana, R., & Niculescu-Mizil, A. (2006). An Empirical Comparison of Supervised Learning Algorithms. Proceedings of the 23rd International Conference on Machine Learning (ICML).
55. Jain, A. K., Murty, M. N., & Flynn, P. J. Data Clustering: A Review. ACM Computing Surveys, 1999; 31(3), 264-323.
56. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
57. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
58. Tan, P. N., Steinbach, M., & Kumar, V. (2018). Introduction to Data Mining (2nd ed.). Pearson.
59. Kaufman, L., &Rousseeuw, P. J. (2009). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley.
60. Everitt, B., Landau, S., Leese, M., & Stahl, D. (2011). Cluster Analysis (5th ed.). Wiley.
61. Jain, A. K. Data Clustering: 50 Years Beyond K-means. Pattern Recognition Letters, 2010; 31(8), 651-666.
62. Ester, M., Kriegel, H. P., Sander, J., & Xu, X.. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), 1996; 226-231.
63. MacQueen, J. Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967; 1(14):281-297
64. Kohonen, T.. The Self-Organizing Map. Proceedings of the IEEE, 1990; 78(9):1464-1480
65. Gramatica, P. Principles of QSAR Models Validation: Internal and External. QSAR & Combinatorial Science, 2007; 26(5), 694-701.
66. Gasteiger, J. (Ed.). (2003). Handbook of Chemoinformatics: From Data to Knowledge (Vol. 4). Wiley-VCH.
67. Chen, H., &Engkvist, O. (2020). Chemoinformatics: Machine Learning in Chemistry. Royal Society of Chemistry.
68. Gramatica, P. (2007). Principles of QSAR Models Validation: Internal and External. QSAR
69. Fourches, D., & Barnes, J. C. (2019). Chemoinformatics and Computational Chemical Biology. Royal Society of Chemistry.
70. Brown, N., & Martin, Y. C. (Eds.). (2017). Computational Chemogenomics. Methods in Pharmacology and Toxicology. Springer.
71. Ramachandran, S., Kundu, S., & Bansal, V. Chemoinformatics Approaches for Virtual Screening and Lead Optimization: Current Scenario in Drug Discovery. Methods, 2012; 57(4):459-468.